e-Portfolio Activity: Reflective Activity 1 – Ethics in Computing in the age of Generative AI

Corrêa et al.'s (2023) review of 200 governance policies and ethical guidelines from around the world is an important research paper as it shone a light on the shortcomings of current ethical policy making, providing a vital dataset and outlining key challenges to this field. Before looking at these challenges, it is interesting to note that the top 5 of 17 common principles referenced within Corrêa et al.'s (2023) analysed ethical papers mirrored those found in Fjeld et al.'s (2020) similar but smaller review of 36 guidelines:

Top 5 shared principles in 200 AI ethics papers (Corrêa et al., 2023):

1. **Transparency/Explainability/Auditability:** (89%).
2. **Justice/Equity/Fairness/Non-discrimination:** (86%).
3. **Accountability/Liability:** (85%).
4. **Privacy:** (84%).
5. **Reliability/Safety/Security/Trustworthiness:** (78%).

Top 5 shared principles in 36 AI ethics papers (Fjeld et al., 2020):

1. **Fairness and Non-discrimination** (100%).
2. **Transparency and Explainability** (94%).
3. **Privacy** (97%).
4. **Accountability** (97%).
5. **Safety and Security** (81%).

Although such commonalities existed, that is not to say that consensus was achieved between each paper; Corrêa et al. (2023) unveiled 5 key challenges to the field of AI ethics:

1. Geographical and cultural biases exist due to underrepresentation of countries;
2. Gender disparity was very apparent: 66% of authors of AI ethics guidelines were male, with only 34% female;

3. Accountability and effectiveness of rulings are not being established well enough, with a severe lack of legally binding regulations being put in place;
4. A disparity exists between what should be happening in terms of AI development or usage and what tools are available to enact change.
5. More focus is needed on the long-term impact of AI as this is not given enough credence in papers.

(Corrêa et al., 2023)

The paper also acknowledged the need for tighter definitions as these often vary between papers – a key issue that is echoed by Fjeld et al.'s (2020) similar review, where, although finding that guidelines on AI ethics tended to prioritise and cover the same key issues, there was a distinct lack of shared definitions and visions. The authors were also in agreeance with issue 4 above, with clear practical application of high-level ethical principles lacking within the guidelines (Fjeld et al.'s, 2020).

In analysing the effectiveness of the Menlo report (Bailey et al., 2012) - a key initiative intent on establishing ethical guidelines for ICT research - Finn & Shilton (2023) found that many consider the project to be a success in that it influenced academic networks of peer review; established a new requirement of ethical statements in top conferences; it influenced how research is viewed and accepted for publication; it drew much needed attention to the field of ethics in ICT research; and it is said to have inspired continual localised ethical review as it uncovered new challenges and unanswered questions. However, similar to a constraining factor for ethical AI guidelines mentioned in Corrêa et al.'s (2023) study, the report did not establish legally binding guidelines and so its uptake was subsequently less widespread as was hoped.

In discussing the attributes of an effective AI ethicist, Deckard (2023) emphasises the importance of having a knowledge of social sciences, excellent communication skills in order to interact with many different civil society organisations, as well as having a commitment to social responsibility. Such social sensitivity appears paramount to addressing the challenges uncovered from the 3 discussed studies within AI ethics policy making; AI is, after all, an imitation and, increasingly, an extension of human cognition (Kurzweil, 2024) suggesting that the human cognition that we choose to extend needs to be just that – a fair representation of all human cognition, not just that of westerners.

Corrêa et al. (2023) promote the use of their dataset and visualisation tool to inform and be extended by further studies in the home hope that the discussed issues may be addressed. In response to these findings, I believe that it is important to map out which countries are currently under-represented within AI policy publications. The focus then should be in establishing a Worldwide AI Ethics committee, with representatives from all major countries - with an equal number of male and female representatives - with the goal of achieving uniformity in AI ethics policy both in terms of establishing policy and in terms of defining terms and concepts. Due to the fast-changing nature of the field, I believe that the committee should meet biannually to review developments. Another primary focus for the committee should be to push for ethics policies to be legally-binding across the world, with an endeavour to drive forward the development of practical tools to enact recommendations. A final focus should be to look at the long-term effects of AI and how ethical guidelines can respond to these.

Such an endeavour can be seen as being ambitious, however the growing impact and accelerating development of AI has been unprecedented, particularly within the last 7 years, and we may soon reach a further point of acceleration – the scale of which is difficult to estimate – if or when AI takes over the development of more advanced AI (Kurzweil, 2024). In my opinion, responding comprehensively to Corrêa et al.'s (2023) findings requires such ambitious action to keep up with this rate of change and oversee the development of AI on humanity's terms. Notwithstanding this, it has to be acknowledged that it would be an extreme challenge to coordinate such a number of committee representatives. It would also prove extremely challenging to establish a consensus of opinion that satisfies all members of the committee. Finally, no doubt the biggest challenge would be to oversee the writing of AI ethics policies into law worldwide.

References

Bailey, M., Dittrich, D., Kenneally, E., & Maughan, D. (2012). 'The Menlo Report'. *IEEE Security & Privacy*, 10(2), pp.71–75. Available at: https://doi.org/10.1109/MSP.2012.52 (Accessed: 01 August 2025).

Corrêa, N.K., Galvão, C., Santos, J.W., Del Pino, C., Pinto, E.P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E. and de Oliveira, N. (2023) 'Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance', *Patterns*, 4(10), pp.1-14. Available at: https://doi.org/10.1016/j.patter.2023.100857 (Accessed: 30 July 2025).

Finn, M. & Shilton, K. (2023) ‘Ethics governance development: The case of the Menlo Report’, *Social Studies of Science*, 53(3), pp. 315–340. Available at: 10.1177/03063127231151708 (Accessed: 29 July 2025).

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., & Srikumar, M. (2020) *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Available at: https://cyber.harvard.edu/publication/2020/principled-ai (Accessed: 31 July 2025).

Kurzweil, R., (2024). *The singularity is nearer: When we merge with AI*. New York: Penguin.